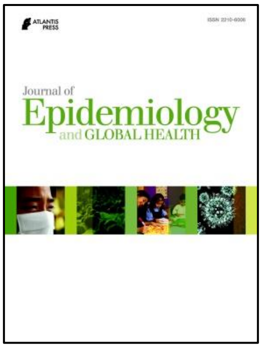

Journal of Epidemiology and Global Health

ISSN (Online): 2210-6014 ISSN (Print): 2210-6006
Journal Home Page: https://www.atlantis-press.com/journals/jegh

Bayesian lead time estimation for the Johns Hopkins Lung Project data

Hyejeong Jang, Seongho Kim, Dongfeng Wu

To cite this article: Hyejeong Jang, Seongho Kim, Dongfeng Wu (2013) Bayesian lead time estimation for the Johns Hopkins Lung Project data, Journal of Epidemiology and Global Health 3:3, 157–163, DOI: https://doi.org/10.1016/j.jegh.2013.05.001

To link to this article: https://doi.org/10.1016/j.jegh.2013.05.001

Published online: 23 April 2019

Journal of Epidemiology and Global Health (2013) 3, 157–163

ELSEVIER

http:// www.elsevier.com/locate/jegh

Bayesian lead time estimation for the Johns Hopkins Lung Project data

Hyejeong Jang [1], Seongho Kim [2], Dongfeng Wu *

Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, KY 40202, United States

Received 13 February 2013; received in revised form 8 May 2013; accepted 10 May 2013
Available online 14 June 2013

KEYWORDS
Lead time;
Lifetime distribution;
X-ray screening;
Lung cancer

Abstract *Problem statement:* Lung cancer screening using X-rays has been controversial for many years. A major concern is whether lung cancer screening really brings any survival benefits, which depends on effective treatment after early detection. The problem was analyzed from a different point of view and estimates were presented of the projected lead time for participants in a lung cancer screening program using the Johns Hopkins Lung Project (JHLP) data.

Method: The newly developed method of lead time estimation was applied where the lifetime T was treated as a random variable rather than a fixed value, resulting in the number of future screenings for a given individual is a random variable. Using the actuarial life table available from the United States Social Security Administration, the lifetime distribution was first obtained, then the lead time distribution was projected using the JHLP data.

Results: The data analysis with the JHLP data shows that, for a male heavy smoker with initial screening ages at 50, 60, and 70, the probability of no-early-detection with semiannual screens will be 32.16%, 32.45%, and 33.17%, respectively; while the mean lead time is 1.36, 1.33 and 1.23 years. The probability of no-early-detection increases monotonically when the screening interval increases, and it increases slightly as the initial age increases for the same screening interval. The mean lead time and its standard error decrease when the screening interval increases for all age groups, and both decrease when initial age increases with the same screening interval.

Conclusion: The overall mean lead time estimated with a random lifetime T is slightly less than that with a fixed value of T. This result is hoped to be of benefit to improve current screening programs.

© 2013 Ministry of Health, Saudi Arabia. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Lung cancer, the leading cause of cancer death for both men and women, occurs in the lungs and most often begins in the cells that line the bronchi. Two

* Corresponding author. Tel.: +1 502 852 1888; fax: +1 502 852 3294.

E-mail addresses: h0jang01@louisville.edu (H. Jang), s0kim023@louisville.edu (S. Kim), dongfeng.wu@louisville.edu (D. Wu).

[1] Tel.: +1 502 852 8078.

[2] Tel.: +1 502 852 3525.

2210-6006/$ - see front matter © 2013 Ministry of Health, Saudi Arabia. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

http://dx.doi.org/10.1016/j.jegh.2013.05.001

major types of lung cancer have been identified: small cell lung cancer, which accounts for about 20% of all cases, and non-small cell lung cancer, the most common type. Different types of lung cancers require different kinds of treatments. The age-specific lung cancer incidence rate rises with advancing age and reaches its peak between 65 and 74 [1].

Cancer screening is carried out to detect malignant tumors early, in order to translate into early treatment and a better prognosis. However, there are controversies concerning lung cancer screening since the early 1970s. The benefit of screening is often measured by collecting information on how long patients are alive after the diagnosis, called survival time. The survival time is the time difference between the date the disease is diagnosed and the date a patient dies due to the disease. However, a patient's survival time is often perceived longer due to the earlier date of diagnosis, even though early detection may not be translated into effective treatment in those days. For example, suppose a screening exam leads to a cancer diagnosis at time t before any symptoms appear, as shown in Fig. 1, then the survival time will be calculated as $(t_d - t)$, although in fact the survival time is $(t_d - t_2)$. This bias occurs due to the contribution of $(t_2 - t)$, which is called the lead time. The lead time is the difference between the time of diagnosis via a screening exam and the time of clinical disease onset without screening [2]. Since the survival benefit of screening heavily relies on the lead time, it is critical to accurately evaluate the distribution of the lead time.

For this study, the commonly followed disease progressive model is assumed where the disease develops by progressing through three states, denoted by $S_0 \rightarrow S_p \rightarrow S_c$ [3]. Its graphical representation is illustrated in Fig. 1. The state S_0 refers to the disease-free state, where either a person does not have the disease, or the disease is in such an early stage that it cannot be detected by a screening exam. The preclinical disease state, S_p, is a state in which an asymptomatic individual unknowingly has the disease that a screening exam can detect. The disease state, S_c, is a state at which the disease manifests itself with clinical symptoms. If a person enters the preclinical state (S_p) at age t_1 and becomes clinically incident (S_c) later at age $t_2(>t_1)$, then $(t_2 - t_1)$ is the sojourn time in the preclinical state. However, if this person undergoes a screening exam at time t within the time interval (t_1, t_2) and cancer is diagnosed, then the length of time $(t_2 - t)$ is the person's lead time.

Many researchers have proposed methods for inference on the lead time among participants in a screening program [4—12], usually by providing formulas to estimate the mean and the variance of the lead time. Wu et al. [2] derived the probability distribution of the lead time for the whole diseased cohort, including both screen-detected cases and interval-incident cases, where the human lifetime was treated as a fixed value. The model allows estimation of the proportion of patients whose lead time is zero and the proportion whose lead time is positive from the program. Later, Wu et al. applied this approach to the Mayo Lung Project (MLP) data to estimate the lead time distribution [13]. However, it is not realistic to fix a person's lifetime T in the estimation of the lead time distribution. For this reason, [18] developed a model to treat the lifetime T as a random variable and made the estimation of lead time distribution more practical [1]. The main objective of the present study is to evaluate the lead time distribution in lung cancer screening using the Johns Hopkins Lung Project (JHLP) data and the newly developed method with a changing lifetime T.

2. Materials and methods

The design of the JHLP can be found in the literature [14]. The JHLP trials enrolled 10,386 men in the Baltimore metropolitan area between 1973 and 1978, aged 45 years and older at enrollment, who smoked at least one pack of cigarettes per day (or who had smoked this much within 1 year of enrollment) and who had no prior history of respiratory cancer. Then all participants were randomized into two groups: chest X-ray only, or a dual screen (chest X-ray and sputum cytology) group, resulting in 5160 men in the chest X-ray only

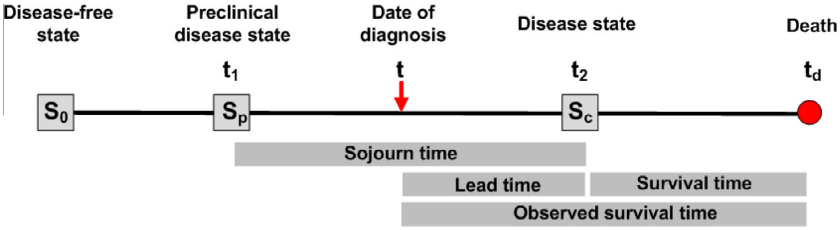

Figure 1 A graphical representation of the disease progressive model.

group and 5226 in the dual-screen group. Participants in the chest X-ray group received a chest X-ray screening test annually for eight consecutive years. If any of the tests was positive, then the screen was considered positive and a definitive work-up exam, such as biopsy, was done. The data that were used in this study were the chest X-ray group, and the data included the number of participants in each screening exam, the number of detected and confirmed cancer cases in each screening exam, and the number of interval cases. These data were stratified by age at study entry from 45 to 88 years old. However, after carefully examining the data, only the data from age 45 to age 70 were used, excluding age groups 47, 58, 62, 68, 69 and ages above 70, because these age groups had very few participants and might cause a large bias in the estimation.

Consider a cohort of initially asymptomatic individuals in a screening program. Let $\beta(t)$ be the sensitivity of the screening modality, where t is the individual's age at the exam. Define $w(t)dt$ as the probability of a transition from S_0 to S_p during (t, $t + dt$). Let $q(x)$ be the probability density function (pdf) of the sojourn time in S_p, and let $Q(z) = \int_z^\infty q(x)dx$ be the survivor function of the sojourn time in the preclinical state S_p. Throughout this paper, the time variable t represents the participating individual's age; the random variable T represents a person's lifetime with a probability density function $f_T(t)$.

For an initially asymptomatic male heavy smoker of age t_0, who has no history of lung cancer, and suppose that he plans to undergo K screening exams at ages $t_0 < t_1 < \ldots < t_{K-1}$. The distribution of the lead time will be a point mass at 0, and a positive continuous probability density. The reason is that for the screen-detected cases, the lead time is greater than zero, while for the interval incident cases, the lead time is zero. A brief summary of the derived probability formulae for the lead time with a changing human lifetime T is given in the Appendix A.

The lead time distribution is a function of the screening sensitivity $\beta(t)$, the transition probability density $w(t)$, the sojourn time distribution $q(x)$, a person's initial age at screening, and a future screening frequency or screening schedule. The first three parameters were estimated from the JHLP data using the following parametric models:

$$\beta(t) = \beta, \tag{1}$$

$$w(t|\mu, \sigma^2) = 0.3 \cdot \exp\left\{-\frac{(\log t - \mu)^2}{2\sigma^2}\right\} \Big/ \left(\sqrt{2\pi}\sigma t\right),$$
$$\sigma > 0, \tag{2}$$

and

$$q(x) = \frac{\kappa x^{\kappa-1}\rho^\kappa}{(1 + x^\kappa \rho^\kappa)^2}, \quad \kappa > 0,\ \rho > 0, \tag{3}$$

where t represents age and x is the sojourn time in the preclinical state S_p. The screening sensitivity was treated as a stable value for all age groups, that is, $\beta(t) = \beta$. The lognormal distribution was chosen for $w(t)$ with an upper limit of 30%. According to the NIH SEER database, the lifetime risk of lung cancer for the general population is about 7% for both genders [15]. Since participants in the JHLP were male heavy smokers, the risk would be higher than that, besides the fact that not all people in the preclinical state will progress into clinical cancer. This research proposes 30% as a reasonable upper limit for $w(t)$. A more detailed description of the parametric models can be found in Wu et al. [13,16].

In this study, the unknown parameters are $\theta = (\beta, \mu, \sigma^2, \kappa, \rho)$. Markov Chain Monte Carlo (MCMC) was used to draw posterior samples with noninformative priors; each simulation was run for 11,000 iterations, with 1000 burn-in steps, and after the burn-in steps, then the posteriors were sampled every 10 steps. The MCMC trace plots and the posterior density plots of θ are plotted using the final 1000 posterior samples for θ, as can be seen in Figs. 2 and 3. All parameters were converged nicely based on Bayesian output analysis. The posterior means of the parameters are $\hat{\theta} = (\hat{\beta}, \hat{\mu}, \hat{\sigma}^2, \hat{\kappa}, \hat{\rho})$ = (0.568, 3.922, 1.020, 1.027, 1.049). Table 1 summarizes the estimates of the parameters.

3. Results

The Bayesian posterior samples θ_i^* in the inference for the lead time were used, where θ_i^* is one of the posterior samples generated from the MCMC. The posterior predictive distribution of the lead-time is:

$$f_L^{\mathrm{JHLP}}(l) \approx \frac{1}{n}\sum_{i=1}^{n} f_L^{\mathrm{JHLP}}(l|\theta_i^*) \tag{4}$$

where θ_i^* is the posterior sample ($i = 1,\ldots,1000$) and $f_L^{HIP}(l|\theta_i^*)$ is the mixture distribution defined by Eqs. (A.5) and (A.6).

For simplicity, three cohorts of initially asymptomatic males were chosen, with initial screening age t_0 = 50, 60, and 70, respectively. For each cohort, various screening frequencies were examined, with screening interval Δ = 6, 12, 18, and 24 months. The number of screenings $K = [(T - t_0)/\Delta]$ is a function of the lifetime T, therefore it is a random variable in the simulation. From Eq. (4), the final distribution of the lead time

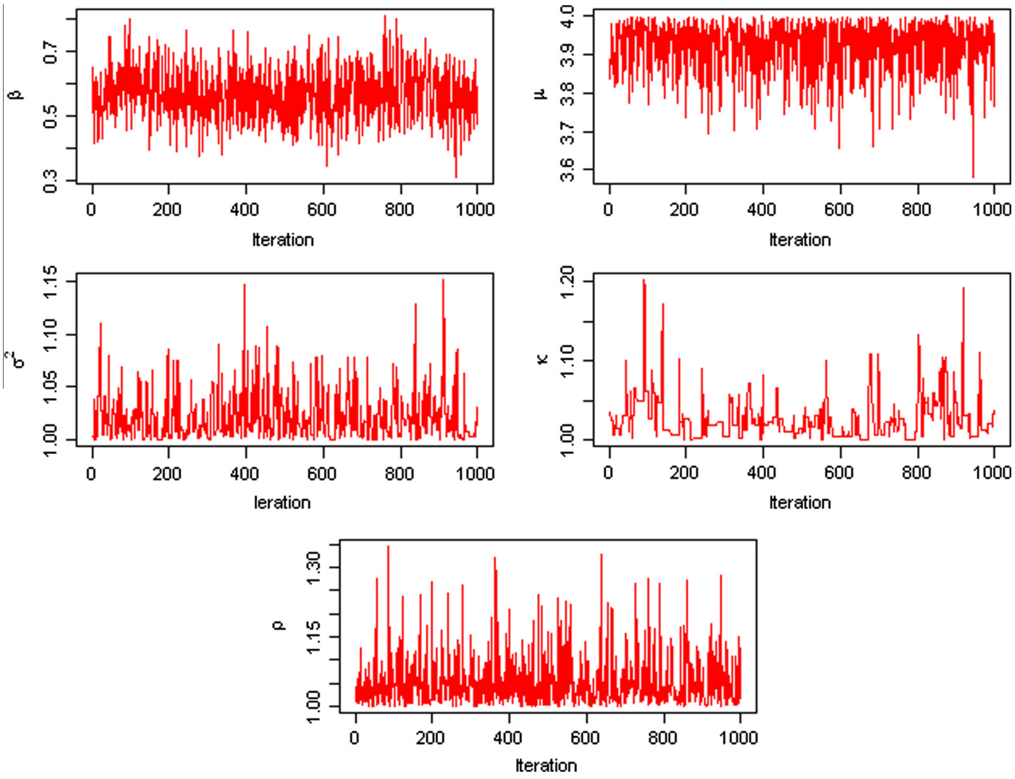

Figure 2 The MCMC trace plots of the parameters $\theta = (\beta, \mu, \sigma^2, \kappa, \rho)$ using the JHLP data.

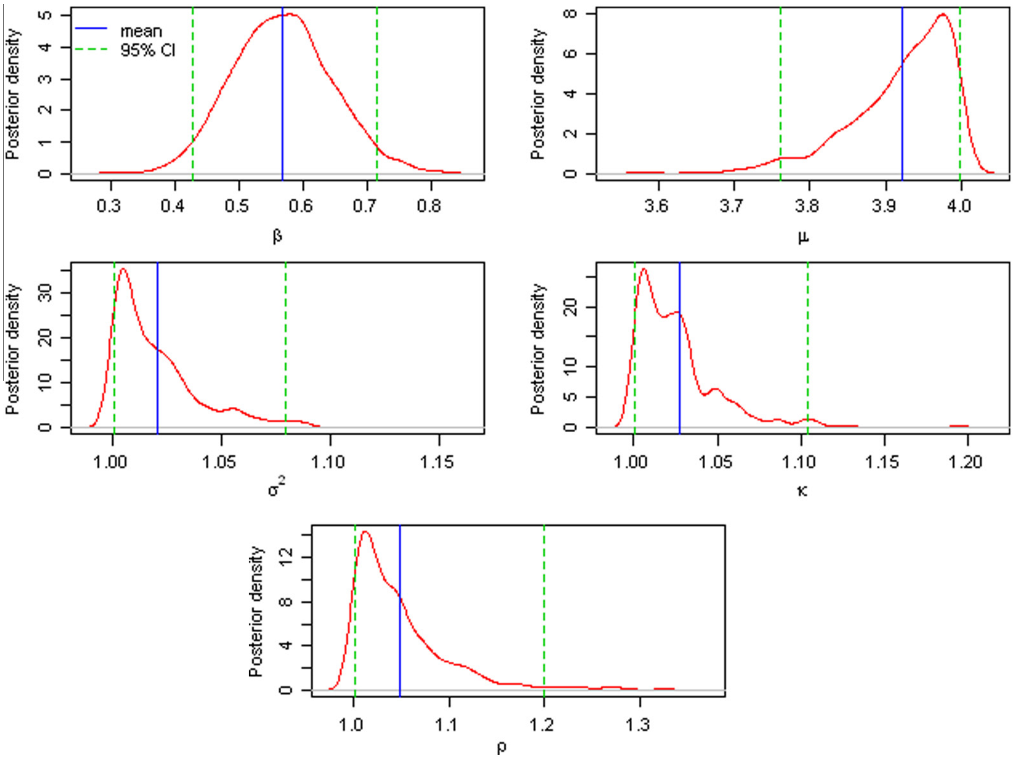

Figure 3 The posterior density plots of the parameters $\theta = (\beta, \mu, \sigma^2, \kappa, \rho)$ using the JHLP data.

Table 1 The estimates of the parameters.

	Mean	SD[a]	2.5%[b]	50%[c]	97.5%[d]
β	0.568	0.076	0.427	0.565	0.716
μ	3.922	0.065	3.762	3.938	3.997
σ^2	1.020	0.022	1.001	1.014	1.079
κ	1.027	0.028	1.001	1.021	1.104
ρ	1.049	0.052	1.001	1.034	1.200

[a] SD stands for the empirical standard deviation.
[b, c, d] The 25th, 50th, and 97.5th percentiles, respectively.

is simply a weighted average of the different lengths of lifetimes.

Table 2 summarizes the Bayesian predictive inference for the lead time. The probability that the lead time is zero and the probability that the lead time is positive are reported as percentages in Table 2. The mean lead time and its empirical standard error were reported in years. The density curves for the lead time are shown in Fig. 4 for different screening intervals when t_0 = 60, as the density curves when the initial screening age is 50 or 70 are very similar.

These results suggest that a man who begins semiannual screening (i.e., Δ = 6 months) when he is 60 years old and develops lung cancer sometime during his life has a 32.45% chance that he will not be detected early by the scheduled screening exams. This probability of no-early-detection from the screening program increases to 46.54% if the exams are annual. For a man with initial screening

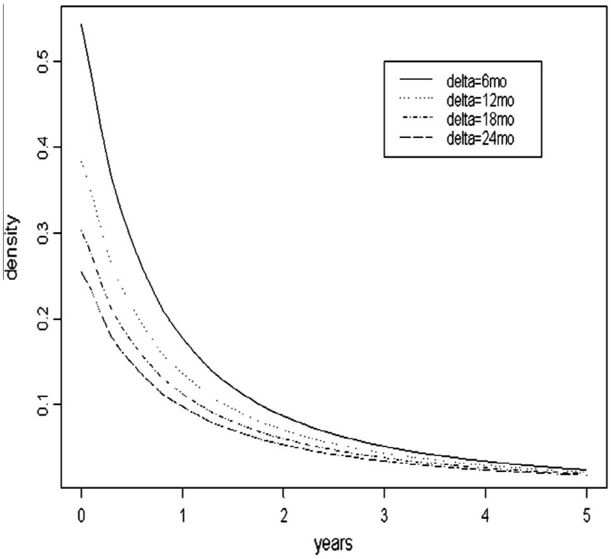

Figure 4 The pdf curve of the lead time when t_0 = 60.

Table 2 A projection of the lead time distribution using the JHLP control group.

Δ[a]	P_0[b] (%)	$1 - P_0$ (%)	Mean (yr)	SE[c] (yr)
Age at initial screen t_0 = 50				
6 mo	32.16	67.84	1.360	2.278
12 mo	46.30	53.70	1.168	2.224
18 mo	54.43	45.57	1.038	2.163
24 mo	59.86	40.14	0.944	2.106
Age at initial screen t_0 = 60				
6 mo	32.45	67.55	1.332	2.229
12 mo	46.54	53.46	1.144	2.175
18 mo	54.58	45.42	1.018	2.116
24 mo	59.97	40.03	0.926	2.060
Age at initial screen t_0 = 70				
6 mo	33.17	66.83	1.230	2.077
12 mo	47.16	52.84	1.051	2.020
18 mo	55.03	44.97	0.933	1.960
24 mo	60.22	39.78	0.848	1.905

[a] $\Delta = t_i - t_{i-1}$ is the time interval between screens.
[b] $P_0 = P(L = 0|D = 1)$ is the probability of ''no-early-detection''.
[c] SE stands for the empirical standard error. This is a simulated projection. The number of screens K is a random variable, changing with the lifetime T.

age at 50 [respectively, 70], the probability of no-early-detection with semiannual screens will be 32.16% [or 33.17% for age 70]. The probability of no-early-detection is monotonically increasing when the screening interval increases within the same age group. This probability is slightly increasing as the initial age increases for the same screening interval. The difference between the initial ages 60 and 70 is smaller than that corresponding difference between the initial age groups 50 and 60.

The mean lead time in each age group decreases as the screening time interval increases in Table 2 (i.e., more frequent screening exams will result in longer lead times). The increase in the mean lead time is due partly to the smaller point mass at zero of the lead time when screening intervals are closer together. The standard deviation of the lead time decreases as the time between screening exams increases. The mean lead time and the standard error of the lead time both slightly decrease as age increases within the same screening interval. Table 2 also reveals that the standard deviation for the lead time is greater than the mean lead time.

4. Discussion

The screening sensitivity, the sojourn time distribution, and the transition density were first estimated in a Bayesian framework. The probability distribution of the lead time from the JHLP data was then estimated by employing a newly developed method [1]. This method considers the human lifetime as a random variable using information from the published actuarial life table of the U.S. Social Security Administration to make the screening model more realistic [17].

The sensitivity was considered as a constant parameter across all age groups in this work. Consequently, the estimated sensitivity $\hat{\beta}$ was 56.8%. Kim et al. (2012) estimated the sensitivity using particle swarm optimization (PSO) using the JHLP study group data, in which both X-ray and sputum cytology were used, when estimating the model parameters. Compared with this previous result, the sensitivity in this study is much smaller than that of the previous result (i.e., 56.8% vs. 79.9%). The reason for the large deviation in sensitivity is that the sensitivity of the previous study was estimated using both X-ray and sputum cytology, while the current study uses the control group data, where only X-ray annual screenings were administered, resulting in a much smaller sensitivity. This confirms that sputum cytology screening improves the overall sensitivity of X-rays.

The density curves of the lead time of the JHLP study group data (i.e., X-ray and cytology screening) were estimated when the lifetime has a fixed upper bound of 75 years old in the previous study [1]. The density curves with the JHLP control group data (i.e., only X-ray) are more skewed to the right than those of the previous study. This in general suggests that the lead time of the JHLP study group data has a greater effect on early detection owing to additional cytology screening, although the lifetime was treated as a random variable for the JHLP control group data.

Some simulation was done for the new lead time model when lifetime T is a random variable; the purpose is to find which factor will affect the distribution of the lead time more significantly. Screening sensitivity was found to affect the lead time distribution slightly; it plays a bigger role in the proportion of no-early-detection versus the proportion of early-detection. The sojourn time plays the most significant role in the lead time distribution: in general, a longer (shorter) sojourn time will lead to a longer (shorter) lead time. For lung cancer, the distribution of the sojourn time is heavily skewed to the right, with a large variance; that is why the variance of the lead time in lung cancer is also large.

The lead time model used in this study can answer the following two main questions: the probability that a person's lung cancer will be detected early if a person would develop lung cancer later in life; and the mean/standard error of the lead time for different screening schedules. It is hoped that the results of this study will be beneficial to improving current screening programs.

Conflict of interest

The authors declare that they have no conflict of interest.

Appendix A.

This appendix provides a summary of the key formulae in the lead time distribution [1]. For an initially asymptomatic male with no history of lung cancer who plans to take K screening exams at ages $t_0 < t_1 < \ldots < t_{K-1}$. Let D represents the true disease status and L represents the lead time. The lead time distribution given that his lifetime $T = t_K (> t_{K-1})$ is:

$$P(L=0|D=1,T=t_K)=P(L=0,D=1|T=t_K)/P(D=1|T=t_K)$$

and

$$f_L(z|D=1,T=t_K)=f_L(z,D=1|T=t_K)/P(D=1|T=t_K), \tag{A.1}$$

where

$$P(D=1|T=t_K) = \int_0^{t_0} w(x)[Q(t_0-x)-Q(t_K-x)]dx + \int_{t_0}^{t_K} w(x)[1-Q(t_K-x)]dx, \quad (A.2)$$

$$P(L=0,D=1|T=t_K)=I_{K,1}+I_{K,2}+\cdots+I_{K,K},$$

$$I_{K,j}=\sum_{i=0}^{j-1}(1-\beta_i)\cdots(1-\beta_{j-1})\int_{t_{i-1}}^{t_i} w(x)[Q(t_{j-1}-x)-Q(t_j-x)]dx + \int_{t_{j-1}}^{t_j} w(x)[1-Q(t_j-x)]dx, \quad \text{for all } j=1,\ldots,K, \quad (A.3)$$

and

$$f_L(z,D=1|T=t_K) = \sum_{i=1}^{j-1}\beta_i\left\{\sum_{r=0}^{i-1}(1-\beta_r)\ldots(1-\beta_{i-1}) \times \int_{t_{r-1}}^{t_r} w(x)q(t_i+z-x)dx + \int_{t_{i-1}}^{t_i} w(x)q(t_i+z-x)dx\right\}I(j>1) + \beta_0\int_0^{t_0} w(x)q(t_0+z-x)dx,$$
$$t_K-t_j<z\leqslant t_K-t_{j-1},\ j=2,3,\ldots,K. \quad (A.4)$$

For an individual currently at age t_0, his lifetime is a random variable, hence the number of screenings is random as well. However, if he plans to follow a pre-planned screening schedule, then the distribution of lead time when the lifetime T is greater than t_0 can be obtained by:

$$P(L=0|D=1,T\geqslant t_0)=\int_{t_0}^{\infty} P(L=0|D=1,T=t)f_T(t|T\geqslant t_0)dt, \quad (A.5)$$

$$f_L(z|D=1,T\geqslant t_0)=\int_{t_0+z}^{\infty} f_L(z|D=1,T=t)f_T(t|T\geqslant t_0)dt, \quad z\in(0,\infty),$$

where the conditional pdf of the lifetime is

$$f_T(t|T\geqslant t_0)=f_T(t)/P(T>t_0) = f_T(t)/[1-F_T(t_0)], \quad \mathit{if}\ t\geqslant t_0. \quad (A.6)$$

This is a valid mixed probability distribution, because it was proved that

$$P(L=0|D=1,T\geqslant t_0)+\int_0^{\infty} f_L(z|D=1,T\geqslant t_0)dz=1.$$

To obtain reliable information on the lifetime distribution, the actuarial life table from the United States Social Security Administration (SSA) was used [17]. For more detailed overviews of the methods for lead time calculation and for the lifetime prediction, refer to Wu et al. (2012).

References

[1] Kim S, Erwin D, Wu D. Efficacy of dual lung cancer screening by chest X-ray and sputum cytology using Johns Hopkins Lung Project data. J Biometrics Biostat 2012;3:3.

[2] Wu D, Rosner GL, Broemeling LD. Bayesian inference for the lead time in periodic cancer screening. Biometrics 2007;63:873—80.

[3] Zelen M, Feinleib M. On the theory of screening for chronic diseases. Biometrika 1969;56:601—14.

[4] Kafadar K, Prorok PC. A data-analytic approach for estimating lead time and screening benefit based on survival curves in randomized cancer screening trials. Stat Med 1994;13:569—86.

[5] Kafadar K, Prorok PC. Computer simulation of randomized cancer screening trials to compare methods of estimating lead time and benefit time. Comput Stat Data Anal 1996;23:263—91.

[6] Kafadar K, Prorok PC. Alternative definitions of comparable case groups and estimates of lead time and benefit time in randomized cancer screening trials. Stat Med 2003;22:83—111.

[7] Kafadar K, Prorok PC, Smith PJ. An estimate of the variance of estimators for lead time and screening benefit time in randomized cancer screening trials. Biometrical J 1998;40: 801—21.

[8] Prorok PC. Bounded recurrence times and lead time in the design of a repetitive screening program. J Appl Prob 1982;19:10—9.

[9] Xu J, Prorok PC. Non-parametric estimation of the post-lead-time survival distribution of screen-detected cancer cases. Stat Med 1995;14:2715—25.

[10] Xu J, Fagerstrom RM, Prorok PC. Estimation of post-lead-time survival under dependence between lead-time and post-lead-time survival. Stat Med 1999;18:155—62.

[11] Walter SD, Day NE. Estimation of the duration of a preclinical disease state using screening data. Am J Epidemiol 1983;118:856—86.

[12] Straatman H, Peer PGM, Verbeek ALM. Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. Biometrics 1997;53:217—29.

[13] Wu D, Erwin D, Rosner GL. Sojourn time and lead time projection in lung cancer screening. Lung Cancer 2011;72: 322—6.

[14] Berlin NI, Buncher CR, Fontana RS, Frost JK, Melamed MR. The national cancer institute cooperative early lung cancer detection program. Results of the initial screen (prevalence). Early lung cancer detection: introduction. Am Rev Respir Dis 1984;130:545—9.

[15] SEER Fast Stats Results, NIH. http://seer.cancer.gov/statfacts/html/lungb.html.

[16] Wu D, Rosner GL, Broemeling LD. MLE and Bayesian inference of age-dependent sensitivity and transition probability in periodic screening. Biometrics 2005;61:1056—63.

[17] Period Life Table. Social Security Administration Actuarial Publications. http://ssa.gov/OACT/STATS/table4c6.html; 2010.

[18] Wu D, Kafadar K, Rosner GL, Broemeling LD. The lead time distribution when lifetime is subject to competing risks in cancer screening. The International Journal of Biostatistics. 2012;8(1):6. http://dx.doi.org/10.1515/1557-4679.1363.

Available online at www.sciencedirect.com

SciVerse ScienceDirect